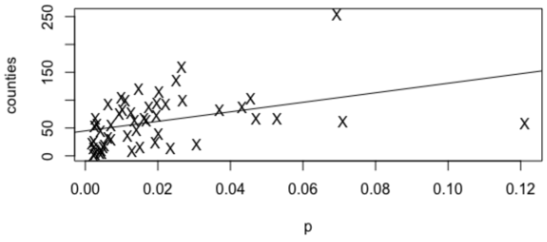
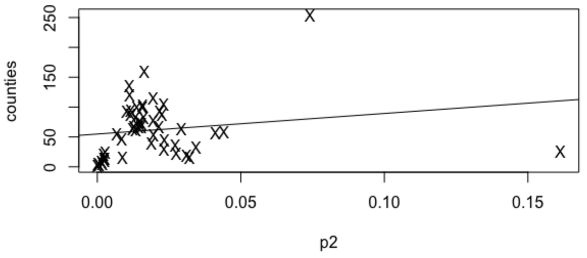
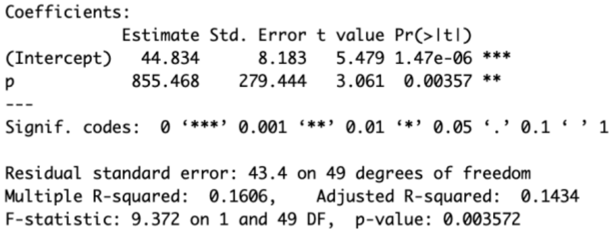
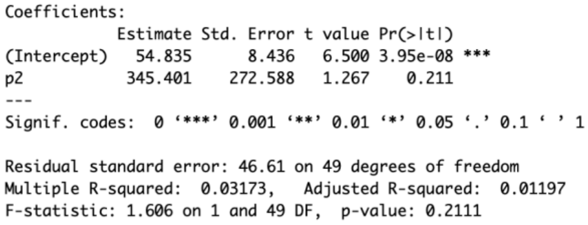
Spencer Palladino
Sampling Test 3

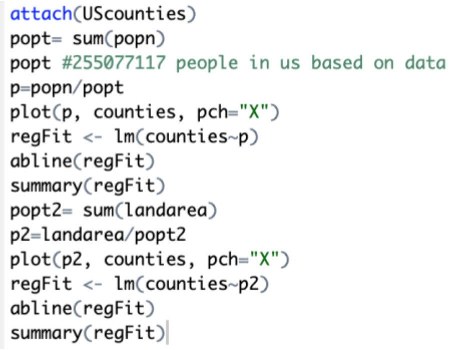
I hereby state that I have not had any form of interaction with any other person in answering this test.

1.

a) I believe that a sample to estimate how many counties are in the US should be selected by using a probability proportional to the population of each state rather than its probability proportional to its area. The reason I believe this is because I found the psi using probability proportional to population is the more correlated to counties. Using the complete data, you can see that when using probability proportional to area that all of the data points are clumped together excluding 2 points. On the other hand, the data seems more linear and correlated when using probability proportional to population. This can also be shown in the p-value for each of the two tests located in the summary below. If we ran a hypothesis test with our null hypothesis being no correlation and our alternative hypothesis being a correlation, we would only be able to reject the null hypothesis (at any reasonable significance level) in the test with probability proportional to the population as it has a p-value of .003.

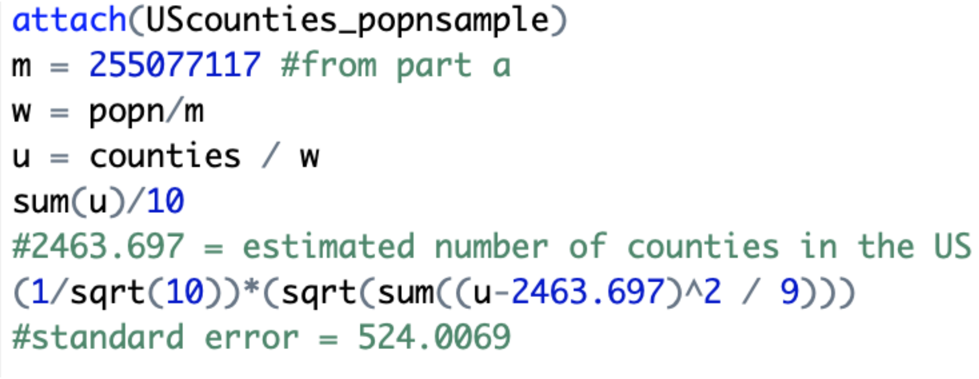
	probability proportional to population	probability proportional to area
graph		
summary	Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 44.834 8.183 5.479 1.47e-06 *** p 855.468 279.444 3.061 0.00357 ** --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 Residual standard error: 43.4 on 49 degrees of freedom Multiple R-squared: 0.1606, Adjusted R-squared: 0.1434 F-statistic: 9.372 on 1 and 49 DF, p-value: 0.003572	Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 54.835 8.436 6.500 3.95e-08 *** p2 345.401 272.588 1.267 0.211 --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 Residual standard error: 46.61 on 49 degrees of freedom Multiple R-squared: 0.03173, Adjusted R-squared: 0.01197 F-statistic: 1.606 on 1 and 49 DF, p-value: 0.2111

Below is the code I used for part a)



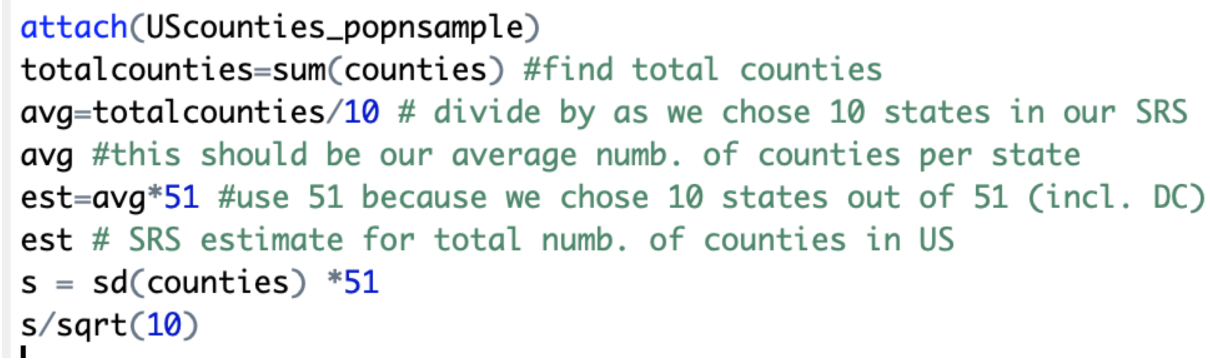
```
attach(UScounties)
popt= sum(popn)
popt #255077117 people in us based on data
p=popn/popt
plot(p, counties, pch="X")
regFit <- lm(counties~p)
abline(regFit)
summary(regFit)
popt2= sum(landarea)
p2=landarea/popt2
plot(p2, counties, pch="X")
regFit <- lm(counties~p2)
abline(regFit)
summary(regFit)
```

b)
Using our sample, I estimated that the total number of counties in the US is = 2463.697 with a standard error of 524.0069 counties.
Below is the code is used for part b)c)



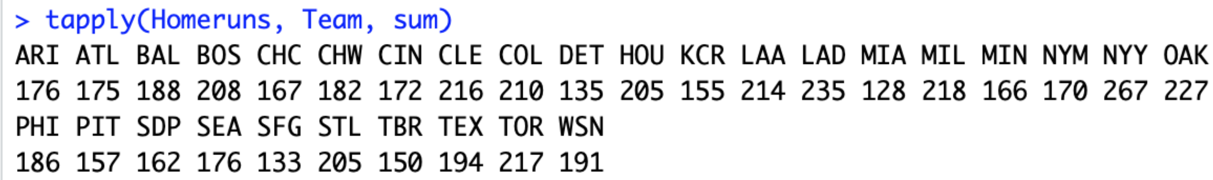
```
attach(UScounties_popnsample)
m = 255077117 #from part a
w = popn/m
u = counties / w
sum(u)/10
#2463.697 = estimated number of counties in the US
(1/sqrt(10))*(sqrt(sum((u-2463.697)^2 / 9)))
#standard error = 524.0069
```

c)
If our friend was just doing a plain SRS, he should have gotten different results than us with an estimated number of counties in the US = 4391.1 and a standard error of 1162.418.
He would have most likely had coded something like this:



```
attach(UScounties_popnsample)
totalcounties=sum(counties) #find total counties
avg=totalcounties/10 # divide by as we chose 10 states in our SRS
avg #this should be our average numb. of counties per state
est=avg*51 #use 51 because we chose 10 states out of 51 (incl. DC)
est # SRS estimate for total numb. of counties in US
s = sd(counties) *51
s/sqrt(10)
```

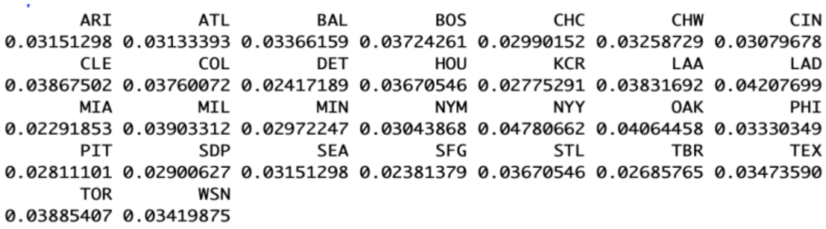
2.
a)



```
> tapply(Homeruns, Team, sum)
ARI ATL BAL BOS CHC CHW CIN CLE COL DET HOU KCR LAA LAD MIA MIL MIN NYM NYY OAK
176 175 188 208 167 182 172 216 210 135 205 155 214 235 128 218 166 170 267 227
PHI PIT SDP SEA SFG STL TBR TEX TOR WSN
186 157 162 176 133 205 150 194 217 191
```

b)
(0.03724261*(0.04207699/(1-0.0372426))) + (0.04207699*(0.0372426/(1-0.04207699)))

Probability that BOS and LAD are selected in a sample size of 2 =
0.003263566



```
       ARI        ATL        BAL        BOS        CHC        CHW        CIN 
0.03151298 0.03133393 0.03366159 0.03724261 0.02990152 0.03258729 0.03079678 
       CLE        COL        DET        HOU        KCR        LAA        LAD 
0.03867502 0.03760072 0.02417189 0.03670546 0.02775291 0.03831692 0.04207699 
       MIA        MIL        MIN        NYM        NYY        OAK        PHI 
0.02291853 0.03903312 0.02972247 0.03043868 0.04780662 0.04064458 0.03330349 
       PIT        SDP        SEA        SFG        STL        TBR        TEX 
0.02811101 0.02900627 0.03151298 0.02381379 0.03670546 0.02685765 0.03473590 
       TOR        WSN 
0.03885407 0.03419875 
```

c)

To find the probability that BOS would be one of the first two teams selected, you would add the probabilities of BOS being selected first and second. Finding the probability that BOS would be selected first is = 0.03724261 which we found from the previous part. The probability that BOS is selected second would be the product of the probability that they failed in the first selection (1-0.03724261) and the probability they were selected second which is dependent on whatever team is selected first.

The equation would look something like this:

P(BOS selected in sample of 2) = 0.03724261 + ((1 - 0.03724261)* (208 / (5585 – x)))

0.03724261 = odds BOS is selected first

208 = BOS Hrs

5585 = total league Hr

x = Hrs of team selected first